



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes

Citation for published version:

ABCTB Investigators & Tomlinson, I 2019, 'Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes', *American Journal of Human Genetics*, vol. 104, no. 1, pp. 21-34.
<https://doi.org/10.1016/j.ajhg.2018.11.002>

Digital Object Identifier (DOI):

[10.1016/j.ajhg.2018.11.002](https://doi.org/10.1016/j.ajhg.2018.11.002)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

American Journal of Human Genetics

Publisher Rights Statement:

This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes

Nasim Mavaddat,^{1,*} Kyriaki Michailidou,^{1,2} Joe Dennis,¹ Michael Lush,¹ Laura Fachal,³ Andrew Lee,¹ Jonathan P. Tyrer,³ Ting-Huei Chen,⁴ Qin Wang,¹ Manjeet K. Bolla,¹ Xin Yang,¹ Muriel A. Adank,⁵ Thomas Ahearn,⁶ Kristiina Aittomäki,⁷ Jamie Allen,¹ Irene L. Andrulis,^{8,9} Hoda Anton-Culver,¹⁰ Natalia N. Antonenkova,¹¹ Volker Arndt,¹² Kristan J. Aronson,¹³ Paul L. Auer,^{14,15} Päivi Auvinen,^{16,17,18} Myrto Barrdahl,¹⁹ Laura E. Beane Freeman,⁶ Matthias W. Beckmann,²⁰ Sabine Behrens,¹⁹ Javier Benitez,^{21,22} Marina Bermisheva,²³ Leslie Bernstein,²⁴ Carl Blomqvist,^{25,26} Natalia V. Bogdanova,^{11,27,28} Stig E. Bojesen,^{29,30,31} Bernardo Bonanni,³² Anne-Lise Børresen-Dale,^{33,34} Hiltrud Brauch,^{35,36,37} Michael Bremer,²⁷ Hermann Brenner,^{12,37,38} Adam Brentnall,³⁹ Ian W. Brock,⁴⁰ Angela Brooks-Wilson,^{41,42} Sara Y. Brucker,⁴³ Thomas Brüning,⁴⁴ Barbara Burwinkel,^{45,46} Daniele Campa,^{19,47} Brian D. Carter,⁴⁸ Jose E. Castelao,⁴⁹ Stephen J. Chanock,⁶ Rowan Chlebowski,⁵⁰ Hans Christiansen,²⁷ Christine L. Clarke,⁵¹ J. Margriet Collée,⁵² Emilie Cordina-Duverger,⁵³ Sten Cornelissen,⁵⁴ Fergus J. Couch,⁵⁵ Angela Cox,⁴⁰ Simon S. Cross,⁵⁶ Kamila Czene,⁵⁷

(Author list continued on next page)

Stratification of women according to their risk of breast cancer based on polygenic risk scores (PRSs) could improve screening and prevention strategies. Our aim was to develop PRSs, optimized for prediction of estrogen receptor (ER)-specific disease, from the largest available genome-wide association dataset and to empirically validate the PRSs in prospective studies. The development dataset comprised 94,075 case subjects and 75,017 control subjects of European ancestry from 69 studies, divided into training and validation sets. Samples were genotyped using genome-wide arrays, and single-nucleotide polymorphisms (SNPs) were selected by stepwise regression or lasso penalized regression. The best performing PRSs were validated in an independent test set comprising 11,428 case subjects and 18,323 control subjects from 10 prospective studies and 190,040 women from UK Biobank (3,215 incident breast cancers). For the best PRSs (313 SNPs), the odds ratio for overall disease per 1 standard deviation in ten prospective studies was 1.61 (95%CI: 1.57–1.65) with area under receiver-operator curve (AUC) = 0.630 (95%CI: 0.628–0.651). The lifetime risk of overall breast cancer in the top centile of the PRSs was 32.6%. Compared with women in the middle quintile, those in the highest 1% of risk had 4.37- and 2.78-fold risks, and those in the lowest 1% of risk had 0.16- and 0.27-fold risks, of developing ER-positive and ER-negative disease, respectively. Goodness-of-fit tests indicated that this PRS was well calibrated and predicts disease risk accurately in the tails of the distribution. This PRS is a powerful and reliable predictor of breast cancer risk that may improve breast cancer prevention programs.

Introduction

Breast cancer is the most common cancer diagnosed among women in Western countries. While rare mutations in genes such as *BRCA1* and *BRCA2* confer high risks of developing

breast cancer, these account for only a small proportion of breast cancer cases in the general population. Multiple common breast cancer susceptibility variants discovered through genome-wide association studies (GWASs)^{1,2} confer small risk individually, but their combined effect, when

¹Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK; ²Department of Electron Microscopy/Molecular Pathology, The Cyprus Institute of Neurology and Genetics, 1683 Nicosia, Cyprus; ³Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK; ⁴Department of Mathematics and Statistics, Laval University, Québec City, QC G1V 0A6, Canada; ⁵Family Cancer Clinic, the Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam, 1066 CX, the Netherlands; ⁶Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, MD 20850, USA; ⁷Department of Clinical Genetics, Helsinki University Hospital, University of Helsinki, Helsinki 00290, Finland; ⁸Fred A. Litwin Center for Cancer Genetics, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada; ⁹Department of Molecular Genetics, University of Toronto, Toronto, ON M5S 1A8, Canada; ¹⁰Department of Epidemiology, Genetic Epidemiology Research Institute, University of California Irvine, Irvine, CA 92617, USA; ¹¹NN Alexandrov Research Institute of Oncology and Medical Radiology, Minsk 223040, Belarus; ¹²Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; ¹³Department of Public Health Sciences, and Cancer Research Institute, Queen's University, Kingston, ON K7L 3N6, Canada; ¹⁴Cancer Prevention Program, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; ¹⁵Zilber School of Public Health, University of Wisconsin-Milwaukee, Milwaukee, WI 53205, USA; ¹⁶Cancer Center, Kuopio University Hospital, Kuopio 70210, Finland; ¹⁷Institute of Clinical Medicine, Oncology, University of Eastern Finland, Kuopio 70210, Finland; ¹⁸Translational Cancer Research Area, University of Eastern Finland, Kuopio 70210, Finland; ¹⁹Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; ²⁰Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen 91054, Germany; ²¹Human Cancer Genetics Programme, Spanish National Cancer Research Centre (CNIO), Madrid 28029, Spain; ²²Biomedical Network on Rare Diseases (CIBERER), Madrid 28029, Spain; ²³Institute of Biochemistry and Genetics, Ufa Scientific Center of Russian Academy of Sciences, Ufa 450054, Russia; ²⁴Department of Population Sciences, Beckman Research Institute of City of Hope, Duarte, CA 91010, USA; ²⁵Department of Oncology, Helsinki University Hospital, University of Helsinki, Helsinki 00290, Finland; ²⁶Department of Oncology, Örebro University Hospital, Örebro 70185, Sweden; ²⁷Department of Radiation Oncology, Hannover Medical

(Affiliations continued on next page)

© 2018 The Authors. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



Mary B. Daly,⁵⁸ Peter Devilee,^{59,60} Thilo Dörk,²⁸ Isabel dos-Santos-Silva,⁶¹ Martine Dumont,⁶² Lorraine Durcan,^{63,64} Miriam Dwek,⁶⁵ Diana M. Eccles,⁶⁴ Arif B. Ekici,⁶⁶ A. Heather Eliassen,^{67,68} Carolina Ellberg,⁶⁹ Christoph Engel,^{70,71} Mikael Eriksson,⁵⁷ D. Gareth Evans,^{72,73} Peter A. Fasching,^{20,74} Jonine Figueroa,^{6,75,76} Olivia Fletcher,⁷⁷ Henrik Flyger,⁷⁸ Asta Försti,^{79,80} Lin Fritschi,⁸¹ Marike Gabrielson,⁵⁷ Manuela Gago-Dominguez,^{82,83} Susan M. Gapstur,⁴⁸ José A. García-Sáenz,⁸⁴ Mia M. Gaudet,⁴⁸ Vassilios Georgoulas,⁸⁵ Graham G. Giles,^{86,87,88} Irina R. Gilyazova,^{23,89} Gord Glendon,⁸ Mark S. Goldberg,^{90,91} David E. Goldgar,⁹² Anna González-Neira,²¹ Grethe I. Grenaker Alnæs,³³ Mervi Grip,⁹³ Jacek Gronwald,⁹⁴ Anne Grundy,⁹⁵ Pascal Guénel,⁵³ Lothar Haeberle,²⁰ Eric Hahnen,^{96,97} Christopher A. Haiman,⁹⁸ Niclas Håkansson,⁹⁹ Ute Hamann,¹⁰⁰ Susan E. Hankinson,¹⁰¹ Elaine F. Harkness,^{102,103,104} Steven N. Hart,¹⁰⁵ Wei He,⁵⁷ Alexander Hein,²⁰ Jane Heyworth,¹⁰⁶ Peter Hillemanns,²⁸ Antoinette Hollestelle,¹⁰⁷ Maartje J. Hooning,¹⁰⁷ Robert N. Hoover,⁶ John L. Hopper,⁸⁷ Anthony Howell,¹⁰⁸ Guanmengqian Huang,¹⁰⁰ Keith Humphreys,⁵⁷ David J. Hunter,^{68,109,110}

(Author list continued on next page)

summarized as a polygenic risk score (PRS), can be substantial.^{3–5} Such genomic profiles can be used to stratify women according to their risk of developing breast cancer.⁶ This in turn holds the promise of improved breast cancer prevention and survival, by targeting screening or other preventative strategies at those women most likely to benefit.

We previously derived a PRS based on 77 established breast cancer susceptibility single-nucleotide polymor-

phisms (SNPs) and reported levels of risk stratification achieved by this PRS.⁷ Based on our findings, several studies have investigated the potential for combining PRSs and other known risk factors for risk stratification and evaluated the impact of risk reduction strategies across risk strata defined by the PRS.^{8–10} Preliminary studies investigating the use of the PRS to inform targeted breast cancer screening programs are underway (see CORDIS

School, Hannover 30625, Germany; ²⁸Gynaecology Research Unit, Hannover Medical School, Hannover 30625, Germany; ²⁹Copenhagen General Population Study, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark; ³⁰Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark; ³¹Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen 2200, Denmark; ³²Division of Cancer Prevention and Genetics, IEO, European Institute of Oncology IRCCS, Milan 20141, Italy; ³³Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital-Radiumhospitalet, Oslo 0379, Norway; ³⁴Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo 0450, Norway; ³⁵Dr Margarete Fischer-Bosch-Institute of Clinical Pharmacology, Stuttgart 70376, Germany; ³⁶University of Tübingen, Tübingen 72074, Germany; ³⁷German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; ³⁸Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg 69120, Germany; ³⁹Centre for Cancer Prevention, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London EC1M 6BQ, UK; ⁴⁰Sheffield Institute for Nucleic Acids (SInFoNIA), Department of Oncology and Metabolism, University of Sheffield, Sheffield S10 2TN, UK; ⁴¹Genome Sciences Centre, BC Cancer Agency, Vancouver, BC V5Z 1L3, Canada; ⁴²Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC V5A 1S6, Canada; ⁴³Department of Gynecology and Obstetrics, University of Tübingen, Tübingen 72076, Germany; ⁴⁴Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum 44789, Germany; ⁴⁵Department of Obstetrics and Gynecology, University of Heidelberg, Heidelberg 69120, Germany; ⁴⁶Molecular Epidemiology Group, C080, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; ⁴⁷Department of Biology, University of Pisa, Pisa 56126, Italy; ⁴⁸Epidemiology Research Program, American Cancer Society, Atlanta, GA 30303, USA; ⁴⁹Oncology and Genetics Unit, Instituto de Investigacion Sanitaria Galicia Sur (IISGS), Xerencia de Xestión Integrada de Vigo-SERGAS, Vigo 36312, Spain; ⁵⁰Division of Medical Oncology and Hematology, University of California at Los Angeles, Los Angeles, CA 90024, USA; ⁵¹Westmead Institute for Medical Research, University of Sydney, Sydney, NSW 2145, Australia; ⁵²Department of Clinical Genetics, Erasmus University Medical Center, Rotterdam 3015 CN, the Netherlands; ⁵³Cancer & Environment Group, Center for Research in Epidemiology and Population Health (CESP), INSERM, University Paris-Sud, University Paris-Saclay, Villejuif 94805, France; ⁵⁴Division of Molecular Pathology, the Netherlands Cancer Institute - Antoni van Leeuwenhoek Hospital, Amsterdam 1066 CX, the Netherlands; ⁵⁵Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA; ⁵⁶Academic Unit of Pathology, Department of Neuroscience, University of Sheffield, Sheffield S10 2TN, UK; ⁵⁷Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 171 65, Sweden; ⁵⁸Department of Clinical Genetics, Fox Chase Cancer Center, Philadelphia, PA 19111, USA; ⁵⁹Department of Pathology, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; ⁶⁰Department of Human Genetics, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; ⁶¹Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK; ⁶²Genomics Center, Centre Hospitalier Universitaire de Québec - Université Laval Research Center, Université Laval, Québec City, QC G1V 4G2, Canada; ⁶³Southampton Clinical Trials Unit, Faculty of Medicine, University of Southampton, Southampton SO17 6YD, UK; ⁶⁴Cancer Sciences Academic Unit, Faculty of Medicine, University of Southampton, Southampton SO17 6YD, UK; ⁶⁵School of Life Sciences, University of Westminster, London W1B 2HW, UK; ⁶⁶Institute of Human Genetics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, Erlangen 91054, Germany; ⁶⁷Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA; ⁶⁸Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, MA 02115, USA; ⁶⁹Department of Cancer Epidemiology, Clinical Sciences, Lund University, Lund 222 42, Sweden; ⁷⁰Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Leipzig 04107, Germany; ⁷¹LIFE - Leipzig Research Centre for Civilization Diseases, University of Leipzig, Leipzig 04103, Germany; ⁷²Division of Evolution and Genomic Sciences, School of Biological Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9WL, UK; ⁷³North West Genomic Laboratory Hub, Manchester Centre for Genomic Medicine, St Mary's Hospital, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9WL, UK; ⁷⁴David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA 90095, USA; ⁷⁵Usher Institute of Population Health Sciences and Informatics, The University of Edinburgh Medical School, Edinburgh EH16 4TJ, UK; ⁷⁶Cancer Research UK Edinburgh Centre, Edinburgh EH4 2XR, UK; ⁷⁷The Breast Cancer Now Toby Robins Research Centre, The Institute of Cancer Research, London SW7 3RP, UK; ⁷⁸Department of Breast Surgery, Herlev and Gentofte Hospital, Copenhagen University Hospital, Herlev 2730, Denmark; ⁷⁹Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; ⁸⁰Center for Primary Health Care Research, Clinical Research Center, Lund University, Malmö 205 02, Sweden; ⁸¹School of Public Health, Curtin University, Perth, WA 6102,

(Affiliations continued on next page)

Milena Jakimovska,¹¹¹ Anna Jakubowska,^{94,112} Wolfgang Janni,¹¹³ Esther M. John,¹¹⁴ Nichola Johnson,⁷⁷ Michael E. Jones,¹¹⁵ Arja Jukkola-Vuorinen,¹¹⁶ Audrey Jung,¹⁹ Rudolf Kaaks,¹⁹ Katarzyna Kaczmarek,⁹⁴ Vesa Kataja,^{18,117} Renske Keeman,⁵⁴ Michael J. Kerin,¹¹⁸ Elza Khusnutdinova,^{23,89} Johanna I. Kiiski,¹¹⁹ Julia A. Knight,^{120,121} Yon-Dschun Ko,¹²² Veli-Matti Kosma,^{18,123,124} Stella Koutros,⁶ Vessela N. Kristensen,^{33,34} Ute Krüger,⁶⁹ Tabea Kühl,¹²⁵ Diether Lambrechts,^{126,127} Loic Le Marchand,¹²⁸ Eunjung Lee,⁹⁸ Flavio Lejbkowitz,¹²⁹ Jenna Lilyquist,¹⁰⁵ Annika Lindblom,¹³⁰ Sara Lindström,^{131,132} Jolanta Lissowska,¹³³ Wing-Yee Lo,^{35,36} Sibylle Loibl,¹³⁴ Jirong Long,¹³⁵ Jan Lubiński,⁹⁴ Michael P. Lux,²⁰ Robert J. MacInnis,^{86,87} Tom Maishman,^{63,64} Enes Makalic,⁸⁷ Ivana Maleva Kostovska,¹¹¹ Arto Mannermaa,^{18,123,124} Siranoush Manoukian,¹³⁶ Sara Margolin,^{137,138} John W.M. Martens,¹⁰⁷ Maria Elena Martinez,^{83,139} Dimitrios Mavroudis,⁸⁵ Catriona McLean,¹⁴⁰ Alfons Meindl,¹⁴¹ Usha Menon,¹⁴² Pooja Middha,^{19,143} Nicola Miller,¹¹⁸ Fernando Moreno,⁸⁴ Anna Marie Mulligan,^{144,145} Claire Mulot,¹⁴⁶

(Author list continued on next page)

and GenomeCanada in [Web Resources](#)).^{11,12} Empirical validation and characterization of the PRS in large-scale epidemiological studies has, however, not been carried out previously. In addition, more informative PRSs would improve the clinical utility of risk prediction. GWASs have now identified ~170 breast cancer susceptibility loci.^{1,2} Moreover, genome-wide heritability estimates indicate that these loci explain only ~40% of the heritability explained by all common variants on genome-wide SNP

arrays. This suggests that the discrimination provided by the PRS could be improved by incorporating variants associated at more liberal significance thresholds. In addition, many variants confer risks that differ by breast cancer subtype (estrogen-receptor [ER]-positive or -negative), suggesting that subtype-specific PRSs might allow better prediction of subtype-specific disease, including the more aggressive ER-negative breast cancer, and enable selection of women for preventative medication.

Australia;⁸²Genomic Medicine Group, Galician Foundation of Genomic Medicine, Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Complejo Hospitalario Universitario de Santiago, SERGAS, Santiago de Compostela 15706, Spain; ⁸³Moore's Cancer Center, University of California San Diego, La Jolla, CA 92093, USA; ⁸⁴Medical Oncology Department, Hospital Clínico San Carlos, Instituto de Investigación Sanitaria San Carlos (IdISSC), Centro Investigación Biomédica en Red de Cáncer (CIBERONC), Madrid 28040, Spain; ⁸⁵Department of Medical Oncology, University Hospital of Heraklion, Heraklion 711 10, Greece; ⁸⁶Cancer Epidemiology & Intelligence Division, Cancer Council Victoria, Melbourne, VIC 3004, Australia; ⁸⁷Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, VIC 3010, Australia; ⁸⁸Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, VIC 3004, Australia; ⁸⁹Department of Genetics and Fundamental Medicine, Bashkir State University, Ufa 450076, Russia; ⁹⁰Department of Medicine, McGill University, Montréal, QC H4A 3J1, Canada; ⁹¹Division of Clinical Epidemiology, Royal Victoria Hospital, McGill University, Montréal, QC H4A 3J1, Canada; ⁹²Department of Dermatology and Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, UT 84112, USA; ⁹³Department of Surgery, Oulu University Hospital, University of Oulu, Oulu 90220, Finland; ⁹⁴Department of Genetics and Pathology, Pomeranian Medical University, Szczecin 71-252, Poland; ⁹⁵Centre de Recherche du Centre Hospitalier de Université de Montréal (CHUM), Université de Montréal, Montréal, QC H2X 0A9, Canada; ⁹⁶Center for Hereditary Breast and Ovarian Cancer, University Hospital of Cologne, Cologne 50937, Germany; ⁹⁷Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne 50931, Germany; ⁹⁸Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA; ⁹⁹Institute of Environmental Medicine, Karolinska Institutet, Stockholm 171 77, Sweden; ¹⁰⁰Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), Heidelberg 69120, Germany; ¹⁰¹Department of Biostatistics & Epidemiology, University of Massachusetts, Amherst, Amherst, MA 1003, USA; ¹⁰²Division of Informatics, Imaging and Data Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester Academic Health Science Centre, Manchester M13 9PT, UK; ¹⁰³Nightingale Breast Screening Centre, Wythenshawe Hospital, Manchester University NHS Foundation Trust, Manchester M23 9LT, UK; ¹⁰⁴NIHR Manchester Biomedical Research Centre, Manchester University NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester M13 9WL, UK; ¹⁰⁵Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA; ¹⁰⁶School of Population and Global Health, University of Western Australia, Perth, WA 6009, Australia; ¹⁰⁷Department of Medical Oncology, Family Cancer Clinic, Erasmus MC Cancer Institute, Rotterdam 3015 CN, the Netherlands; ¹⁰⁸Division of Cancer Sciences, University of Manchester, Manchester M13 9PL, UK; ¹⁰⁹Program in Genetic Epidemiology and Statistical Genetics, Harvard TH Chan School of Public Health, Boston, MA 02115, USA; ¹¹⁰Nuffield Department of Population Health, University of Oxford, Oxford OX3 7LE, UK; ¹¹¹Research Centre for Genetic Engineering and Biotechnology "Georgi D. Efremov," Macedonian Academy of Sciences and Arts, Skopje 1000, Republic of Macedonia; ¹¹²Independent Laboratory of Molecular Biology and Genetic Diagnostics, Pomeranian Medical University, Szczecin 71-252, Poland; ¹¹³Department of Gynecology and Obstetrics, University Hospital Ulm, Ulm 89075, Germany; ¹¹⁴Department of Medicine, Division of Oncology, Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94304, USA; ¹¹⁵Division of Genetics and Epidemiology, The Institute of Cancer Research, London SM2 5NG, UK; ¹¹⁶Department of Oncology, Tampere University Hospital, Tampere, Finland Box 2000, 33521 Tampere, Finland; ¹¹⁷Central Finland Health Care District, Jyväskylä Central Hospital, Jyväskylä 40620, Finland; ¹¹⁸Surgery, School of Medicine, National University of Ireland, Galway H91TK33, Ireland; ¹¹⁹Department of Obstetrics and Gynecology, Helsinki University Hospital, University of Helsinki, Helsinki 00290, Finland; ¹²⁰Prosserman Centre for Population Health Research, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, ON M5T 3L9, Canada; ¹²¹Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON M5T 3M7, Canada; ¹²²Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn 53177, Germany; ¹²³Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, Kuopio 70210, Finland; ¹²⁴Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, Kuopio 70210, Finland; ¹²⁵Cancer Epidemiology Group, University Cancer Center Hamburg (UCCH), University Medical Center Hamburg-Eppendorf, Hamburg 20246, Germany; ¹²⁶VIB Center for Cancer Biology, VIB, Leuven 3000, Belgium; ¹²⁷Laboratory for Translational Genetics, Department of Human Genetics, University of Leuven, Leuven 3000, Belgium; ¹²⁸Epidemiology Program, University of Hawaii Cancer Center, Honolulu, HI 96813, USA; ¹²⁹Clalit National Cancer Control Center, Carmel Medical Center and Technion Faculty of Medicine, Haifa 35254, Israel; ¹³⁰Department of Molecular Medicine and Surgery, Karolinska Institutet, and Department of Clinical Genetics, Karolinska University Hospital, Stockholm 171 76, Sweden; ¹³¹Department of Epidemiology, University of Washington School of Public Health, Seattle, WA 98195, USA; ¹³²Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA; ¹³³Department of Cancer Epidemiology and Prevention, M Skłodowska-Curie Cancer Center - Oncology Institute, Warsaw 02-034, Poland; ¹³⁴German Breast Group, GmbH, Neu Isenburg 63263, Germany; ¹³⁵Division of Epidemiology, Department

(Affiliations continued on next page)

Victor M. Muñoz-Garzon,¹⁴⁷ Susan L. Neuhausen,²⁴ Heli Nevanlinna,¹¹⁹ Patrick Neven,¹⁴⁸ William G. Newman,^{72,73} Sune F. Nielsen,^{29,30} Børge G. Nordestgaard,^{29,30,31} Aaron Norman,¹⁰⁵ Kenneth Offit,^{149,150} Janet E. Olson,¹⁰⁵ Håkan Olsson,⁶⁹ Nick Orr,¹⁵¹ V. Shane Pankratz,¹⁵² Tjong-Won Park-Simon,²⁸ Jose I.A. Perez,¹⁵³ Clara Pérez-Barrios,¹⁵⁴ Paolo Peterlongo,¹⁵⁵ Julian Peto,⁶¹ Mila Pinchev,¹²⁹ Dijana Plaseska-Karanfilska,¹¹¹ Eric C. Polley,¹⁰⁵ Ross Prentice,¹⁴ Nadege Presneau,⁶⁵ Darya Prokofyeva,⁸⁹ Kristen Purrington,¹⁵⁶ Katri Pyrkäs,^{157,158} Brigitte Rack,¹¹³ Paolo Radice,¹⁵⁹ Rohini Rau-Murthy,¹⁵⁰ Gad Rennert,¹²⁹ Hedy S. Rennert,¹²⁹ Valerie Rhenius,³ Mark Robson,¹⁵⁰ Atocha Romero,¹⁵⁴ Kathryn J. Ruddy,¹⁶⁰ Matthias Ruebner,²⁰ Emmanouil Saloustros,¹⁶¹ Dale P. Sandler,¹⁶² Elinor J. Sawyer,¹⁶³ Daniel F. Schmidt,^{87,164} Rita K. Schmutzler,^{96,97} Andreas Schneeweiss,¹⁶⁵ Minouk J. Schoemaker,¹¹⁵ Fredrick Schumacher,¹⁶⁶ Peter Schürmann,²⁸ Lukas Schwentner,¹¹³ Christopher Scott,¹⁰⁵ Rodney J. Scott,^{167,168,169} Caroline Seynaeve,¹⁰⁷ Mitul Shah,³ Mark E. Sherman,¹⁷⁰

(Author list continued on next page)

Here, we used data from 79 studies conducted by the Breast Cancer Association Consortium (BCAC) to optimize PRSs for overall and subtype-specific disease, and we validate their performance in independent datasets.^{1,13–15}

Material and Methods

Study Subjects and Genotyping

The dataset used for development of the PRSs comprised 94,075 breast cancer-affected case subjects and 75,017 control subjects

of European ancestry from 69 studies in the BCAC (Tables S1 and S2). Data collection for individual studies is described previously.¹ Samples were genotyped using one of two arrays: iCOGS^{13,14} and OncoArray.^{1,15} The dataset was divided into a training and validation set. The validation set was randomly selected (approximately 10% of case and control subjects) from studies that had been genotyped with the OncoArray, after excluding studies of bilateral breast cancer, studies or sub-studies oversampling for family history, and individuals with *in situ* cancers or case subjects with unknown ER status.

The best PRSs were evaluated in an independent test dataset comprising 11,428 invasive breast cancer-affected case subjects

of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37232, USA; ¹³⁶Unit of Medical Genetics, Department of Medical Oncology and Hematology, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan 20133, Italy; ¹³⁷Department of Clinical Science and Education, Södersjukhuset, Karolinska Institutet, Stockholm 118 83, Sweden; ¹³⁸Department of Oncology, Södersjukhuset, Stockholm 118 83, Sweden; ¹³⁹Department of Family Medicine and Public Health, University of California San Diego, La Jolla, CA 92093, USA; ¹⁴⁰Anatomical Pathology, The Alfred Hospital, Melbourne, VIC 3004, Australia; ¹⁴¹Department of Gynecology and Obstetrics, Ludwig Maximilian University of Munich, Munich 80336, Germany; ¹⁴²MRC Clinical Trials Unit at UCL, Institute of Clinical Trials & Methodology, University College London, London WC1V 6LJ, UK; ¹⁴³Faculty of Medicine, University of Heidelberg, Heidelberg 69120, Germany; ¹⁴⁴Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON M5S 1A8, Canada; ¹⁴⁵Laboratory Medicine Program, University Health Network, Toronto, ON M5G 2C4, Canada; ¹⁴⁶Université Paris Sorbonne Cité, INSERM UMR-S1147, Paris 75270, France; ¹⁴⁷Radiation Oncology, Hospital Meixoeiro-XXI de Vigo, Vigo 36214, Spain; ¹⁴⁸Leuven Multidisciplinary Breast Center, Department of Oncology, Leuven Cancer Institute, University Hospitals Leuven, Leuven 3000, Belgium; ¹⁴⁹Clinical Genetics Research Lab, Department of Cancer Biology and Genetics, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; ¹⁵⁰Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA; ¹⁵¹Centre for Cancer Research and Cell Biology, Queen's University Belfast, Belfast BT7 1NN, UK; ¹⁵²University of New Mexico Health Sciences Center, University of New Mexico, Albuquerque, NM 87131, USA; ¹⁵³Servicio de Cirugía General y Especialidades, Hospital Monte Naranco, Oviedo 33012, Spain; ¹⁵⁴Medical Oncology Department, Hospital Universitario Puerta de Hierro, Madrid 28222, Spain; ¹⁵⁵Genome Diagnostic Program, IFOM the FIRC (Italian Foundation for Cancer Research) Institute of Molecular Oncology, Milan 20139, Italy; ¹⁵⁶Department of Oncology, Wayne State University School of Medicine, Detroit, MI 48201, USA; ¹⁵⁷Laboratory of Cancer Genetics and Tumor Biology, Cancer and Translational Medicine Research Unit, Biocenter Oulu, University of Oulu, Oulu 90220, Finland; ¹⁵⁸Laboratory of Cancer Genetics and Tumor Biology, Northern Finland Laboratory Centre Oulu, Oulu 90220, Finland; ¹⁵⁹Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Research, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), Milan 20133, Italy; ¹⁶⁰Department of Oncology, Mayo Clinic, Rochester, MN 55905, USA; ¹⁶¹Department of Oncology, University Hospital of Larissa, Larissa 711 10, Greece; ¹⁶²Epidemiology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA; ¹⁶³Research Oncology, Guy's Hospital, King's College London, London SE1 9RT, UK; ¹⁶⁴Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia; ¹⁶⁵National Center for Tumor Diseases, University Hospital and German Cancer Research Center, Heidelberg 69120, Germany; ¹⁶⁶Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA; ¹⁶⁷Division of Molecular Medicine, Pathology North, John Hunter Hospital, Newcastle, NSW 2305, Australia; ¹⁶⁸Discipline of Medical Genetics, School of Biomedical Sciences and Pharmacy, Faculty of Health, University of Newcastle, Callaghan, NSW 2308, Australia; ¹⁶⁹Hunter Medical Research Institute, John Hunter Hospital, Newcastle, NSW 2305, Australia; ¹⁷⁰Department of Health Sciences Research, Mayo Clinic College of Medicine, Jacksonville, FL 32224, USA; ¹⁷¹Precision Medicine, School of Clinical Sciences at Monash Health, Monash University, Clayton, VIC 3168, Australia; ¹⁷²Department of Clinical Pathology, The University of Melbourne, Melbourne, VIC 3010, Australia; ¹⁷³Population Oncology, BC Cancer, Vancouver, BC V5Z 1G1, Canada; ¹⁷⁴School of Population and Public Health, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada; ¹⁷⁵Saarland Cancer Registry, Saarbrücken 66119, Germany; ¹⁷⁶The Curtin UWA Centre for Genetic Origins of Health and Disease, Curtin University and University of Western Australia, Perth, WA 6000, Australia; ¹⁷⁷Division of Breast Cancer Research, The Institute of Cancer Research, London SW7 3RP, UK; ¹⁷⁸Faculty of Medicine, University of Southampton, Southampton SO17 1BJ, UK; ¹⁷⁹Epigenetic and Stem Cell Biology Laboratory, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA; ¹⁸⁰Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY 10032, USA; ¹⁸¹Department of Surgery, Leiden University Medical Center, Leiden 2333 ZA, the Netherlands; ¹⁸²Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham B15 2TT, UK; ¹⁸³Wellcome Trust Centre for Human Genetics and Oxford NIHR Biomedical Research Centre, University of Oxford, Oxford OX3 7BN, UK; ¹⁸⁴Department of Pathology, University Hospital of Heraklion, Heraklion 711 10, Greece; ¹⁸⁵Frauenklinik der Stadtklinik Baden-Baden, Baden-Baden 76532, Germany; ¹⁸⁶Department of Gynecology and Obstetrics, Helios Clinics Berlin-Buch, Berlin 13125, Germany; ¹⁸⁷Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences, NIH, Research Triangle Park, NC 27709, USA; ¹⁸⁸Department of Health Research and Policy - Epidemiology, Stanford

(Affiliations continued on next page)

Martha J. Shrubsole,¹³⁵ Xiao-Ou Shu,¹³⁵ Susan Slager,¹⁰⁵ Ann Smeets,¹⁴⁸ Christof Sohn,¹⁶⁵ Penny Soucy,⁶² Melissa C. Southey,^{171,172} John J. Spinelli,^{173,174} Christa Stegmaier,¹⁷⁵ Jennifer Stone,^{87,176} Anthony J. Swerdlow,^{115,177} Rulla M. Tamimi,^{67,68,109} William J. Tapper,¹⁷⁸ Jack A. Taylor,^{162,179} Mary Beth Terry,¹⁸⁰ Kathrin Thöne,¹²⁵ Rob A.E.M. Tollenaar,¹⁸¹ Ian Tomlinson,^{182,183} Thérèse Truong,⁵³ Maria Tzardi,¹⁸⁴ Hans-Ulrich Ulmer,¹⁸⁵ Michael Untch,¹⁸⁶ Celine M. Vachon,¹⁰⁵ Elke M. van Veen,^{72,73} Joseph Vijai,^{149,150} Clarice R. Weinberg,¹⁸⁷ Camilla Wendt,^{137,138} Alice S. Whittemore,^{188,189} Hans Wildiers,¹⁴⁸ Walter Willett,^{68,190,191} Robert Winqvist,^{157,158} Alicja Wolk,^{99,192} Xiaohong R. Yang,⁶ Drakoulis Yannoukakos,¹⁹³ Yan Zhang,¹² Wei Zheng,¹³⁵ Argyrios Ziogas,¹⁰ ABCTB Investigators,¹⁹⁴ kConFab/AOCS Investigators,¹⁹⁵ NBCS Collaborators,^{33,34,196,197,198,199,200,201,202,203,204,205} Alison M. Dunning,³ Deborah J. Thompson,¹ Georgia Chenevix-Trench,²⁰⁶ Jenny Chang-Claude,^{19,125} Marjanka K. Schmidt,^{54,207} Per Hall,^{57,138} Roger L. Milne,^{86,87,171} Paul D.P. Pharoah,^{1,3} Antonis C. Antoniou,¹ Nilanjan Chatterjee,^{6,208,209} Peter Kraft,^{68,109} Montserrat García-Closas,⁶ Jacques Simard,⁶² and Douglas F. Easton^{1,3}

and 18,323 control subjects from ten studies nested within prospective cohorts, all genotyped using the OncoArray (Tables S3 and S4). The overall breast cancer PRS was also evaluated among 190,040 women of European ancestry from the UK Biobank cohort who had not had any cancer diagnosis or mastectomy prior to recruitment. A total of 3,215 incident registry-confirmed invasive breast cancers developed over 1,381,019 person years of prospective follow-up. Follow-up started 6 months after age of baseline questionnaire. The primary endpoint was invasive breast cancer. Follow-up was censored at the earliest of: risk-reducing mastectomy, diagnosis of any type of cancer, death, or January 15, 2017.

Genotype calling, quality control, and imputation for iCOGS and OncoArray were performed as previously described.^{1,14} Briefly, imputation was performed for the iCOGS and OncoArray datasets separately using the Phase 3 (October 2014) release of the 1000 Genomes data as reference.¹⁶ We followed a two-stage approach using SHAPEIT for phasing¹⁷ and IMPUTE2 for the imputation.¹⁵ Where samples were genotyped with iCOGS and OncoArray, the OncoArray calling was used. SNPs with MAF > 0.01 and imputation $r^2 > 0.9$ for OncoArray and $r^2 > 0.3$ for iCOGS were included in this analysis (~7 million SNPs); a higher threshold was imposed for OncoArray to ensure accurate determination of the PRS in the validation and test datasets.

UK Biobank samples were genotyped using Affymetrix UK BiLEVE Axiom array and Affymetrix UK Biobank Axiom array and imputed to the combined 1000 Genomes Project v.3 and UK10K reference panels using SHAPEIT3 and IMPUTE3.¹⁸ The lowest imputation info score for the SNPs used in these analyses

was 0.86. Samples were included on the basis of female sex (genetic and self-reported) and ethnicity filter (Europeans/White British ancestry subset). Duplicates, individuals with high degree of relatedness (>10 relatives), and one of each related pair of first degree relatives were removed. Samples were also excluded using standard quality control criteria.

Participants provided written informed consent, all studies were approved by the relevant ethics committees, and procedures followed were in accordance with the ethical standards of these committees.

Statistical Analysis

The general aim was to derive a PRS of the form:

$$\text{PRS} = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k \dots + \beta_n x_n$$

where β_k is the per-allele log odds ratio (OR) for breast cancer associated with SNP k , x_k is the allele dosage for SNP k , and n is the total number of SNPs included in the PRS. Previous analyses found no evidence for statistically significant interactions between SNPs^{19,20} and little evidence for departures from a log-additive model for individual SNPs. Assuming this is true in general, the PRS summarizes efficiently the combined effects of SNPs on disease risk.

The main challenge is how to determine which SNPs to include and the weighting parameters β_k to assign. Inclusion of only those SNPs reaching a stringent significance threshold ("genome-wide significant," $p < 5 \times 10^{-8}$) threshold ignores information from larger numbers of SNPs that are likely, but not certain, to be associated with the risk of breast cancer. We used two general

University School of Medicine, Stanford, CA 94305, USA; ¹⁸⁹Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA 94305, USA; ¹⁹⁰Department of Nutrition, Harvard TH Chan School of Public Health, Boston, MA 02115, USA; ¹⁹¹Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA; ¹⁹²Department of Surgical Sciences, Uppsala University, Uppsala 751 05, Sweden; ¹⁹³Molecular Diagnostics Laboratory, INRASTES, National Centre for Scientific Research "Demokritos," Athens 15310, Greece; ¹⁹⁴Australian Breast Cancer Tissue Bank, Westmead Institute for Medical Research, University of Sydney, Sydney, NSW 2145, Australia; ¹⁹⁵Peter MacCallum Cancer Center, Melbourne, VIC 3000, Australia; ¹⁹⁶Department of Research, Vestre Viken Hospital, Drammen 3019, Norway; ¹⁹⁷Department of Cancer Genetics, Vestre Viken Hospital, Drammen 3019, Norway; ¹⁹⁸Section for Breast and Endocrine Surgery, Department of Cancer, Division of Surgery, Cancer and Transplantation Medicine, Oslo University Hospital-Ullevål, Oslo 0450, Norway; ¹⁹⁹Department of Radiology and Nuclear Medicine, Oslo University Hospital, Oslo 0379, Norway; ²⁰⁰Department of Pathology, Akershus University Hospital, Lørenskog 1478, Norway; ²⁰¹Department of Tumor Biology, Institute for Cancer Research, Oslo University Hospital, Oslo 0379, Norway; ²⁰²Department of Oncology, Division of Surgery, Cancer and Transplantation Medicine, Oslo University Hospital-Radiumhospitalet, Oslo 0379, Norway; ²⁰³National Advisory Unit on Late Effects after Cancer Treatment, Oslo University Hospital-Radiumhospitalet, Oslo 0379, Norway; ²⁰⁴Department of Oncology, Akershus University Hospital, Lørenskog 1478, Norway; ²⁰⁵Breast Cancer Research Consortium, Oslo University Hospital, Oslo 0379, Norway; ²⁰⁶Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia; ²⁰⁷Division of Psychosocial Research and Epidemiology, the Netherlands Cancer Institute - Antoni van Leeuwenhoek hospital, Amsterdam 1066 CX, the Netherlands; ²⁰⁸Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA; ²⁰⁹Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA

*Correspondence: nm274@medschl.cam.ac.uk

<https://doi.org/10.1016/j.ajhg.2018.11.002>

approaches for model selection: “hard-thresholding,” based on a stepwise regression model that retained SNPs significantly associated with overall or subtype-specific disease at a given threshold, and penalized regression using lasso.^{21,22} A schema for the analyses is shown in [Figure S1](#).

To prioritize SNPs for analysis, single SNP association tests were first conducted in the training set. Per-allele ORs and standard errors were estimated separately in the iCOGS and OncoArray datasets, adjusting for study and nine ancestry informative principal components (PCs) in the iCOGS dataset and by country and ten PCs in the OncoArray dataset, using a purpose-written program.¹ Combined *p* values were then derived using a fixed-effects meta-analysis with the software METAL.²³ SNPs were sorted by *p* value and filtered on LD, such that uncorrelated SNPs (correlation $r^2 < 0.9$) with lowest *p* value for association with overall breast cancer in the training set were retained (more rigorous pruning, for example at $r^2 < 0.2$, would have removed from consideration informative SNPs from regions with multiple correlated signals^{24,25}).

In the hard thresholding approach, a series of stepwise forward regression analyses were first carried out in 1 Mb regions centered on SNPs significant at a pre-specified threshold for association with either overall and/or subtype-specific disease in the training set. Only SNPs passing the specified *p* value thresholds were included in each 1 Mb region. Two analyses were performed in parallel: for overall breast cancer and ER-negative disease. At each stage the SNP with the smallest (conditional) *p* value for any analysis was added to the model, the threshold for the stepwise regression being the same as that for pre-selection. The process was repeated until no further SNPs could be added at the pre-defined threshold. A second stage of stepwise regressions were then carried out across all regions in each chromosome, to take into account correlated SNPs in different regions. Finally, the effect sizes for the selected SNPs were jointly estimated in a single logistic regression model.

For the best-performing PRSs, SNPs associated with ER-positive at $p < 10^{-6}$ but not with overall breast cancer (at $p < 10^{-5}$) were added at the end of the final SNP list. A third round of stepwise forward regression was then carried out with *p* value for selection of $p < 10^{-6}$ for ER-positive disease. For completeness we added to this final PRS two rarer variants (*BRCA2* p.Lys3326X and *CHEK2* p.Ile157Tyr) which are established to confer a moderate risk of breast cancer and were genotyped on the OncoArray but did not pass the allele frequency threshold in the PRS development phase.

For the penalized regression using lasso, we used the program *glmnet*²¹. SNPs with $p < 0.001$ in overall BC or ER-negative disease in the training set were pre-selected for inclusion in the lasso, and *BRCA2* p.Lys3326X and *CHEK2* p.Ile157Thr were added. Covariates for 19 PCs (9 for iCOGs and 10 for Oncoarray) and country were included in each model. For overall breast cancer, the penalty parameter (lambda) giving the best overall breast cancer PRS in the validation set was selected.

To construct subtype-specific PRSs, we evaluated four different methods: (1) using effect sizes for overall breast cancer (for each of the subtypes), (2) using effect sizes for subtype-specific (ER-positive or ER-negative) disease, (3) using a hybrid method, in which effect sizes were estimated in the relevant subtype for SNPs passing a certain optimal significance threshold in a case-only logistic regression (ER-positive versus ER-negative disease), and otherwise, using effect sizes estimated for overall breast cancer, or (4) by estimating case-only ORs using lasso and combining these with the

overall breast cancer ORs to derive subtype-specific estimates, using the formulae:

$$\beta_{\text{ERpositive}} = \beta_{\text{overall}} + \eta * \beta_{\text{case-only}}$$

$$\beta_{\text{ERnegative}} = \beta_{\text{overall}} - (1 - \eta) * \beta_{\text{case-only}}$$

where $\eta = 0.27$ was the proportion of ER-negative tumors in the validation set.

For the lasso analysis, effect sizes for subtype-specific disease were estimated using method 4 above, combining the estimates from a case-only lasso analysis with the coefficients for overall breast cancer from the lasso analysis. The lambda for the case-only model giving the best subtype-specific PRS in the validation set was selected.

To evaluate the performance of each potential PRS, we standardized the PRSs to have unit standard deviation (SD) in the validation set of control subjects. The association of the standardized PRSs was evaluated in the validation and test (prospective studies) datasets, by logistic regression. We used a Cox proportional hazards regression model to assess the association with risk of breast cancer in UK Biobank. Models were also compared in terms of the area under the receiver operator characteristic curves (AUC), adjusted for study, calculated using the Stata command *comproc*. Meta-analysis of study-specific effects was carried out using the Stata command *metan*.

The goodness of fit of the continuous model (i.e., assuming a linear association between log(OR) and risk) was tested using the Hosmer-Lemeshow (HL) test to compare the observed and predicted risks by quantile and using the tail-based test proposed by Song et al.²⁶ In addition, we considered specifically the risks in the highest and lowest 1% of the distribution.

Effect modification of the PRS by age and family history of breast cancer in first-degree relatives was evaluated by fitting additional interaction terms in the model. The validation and prospective test datasets were combined for this analysis.

The absolute risks of developing breast cancer (overall and subtype-specific disease) were calculated taking into account the competing risk of dying from causes other than breast cancer, as described previously,⁷ with the PRS modeled as a continuous covariate and including a linear “age × PRS” interaction term. The absolute risk of developing subtype-specific disease was obtained constraining to the incidence of overall incidence of ER-negative and ER-positive disease in the UK. Women are at risk of developing both ER-negative and ER-positive disease, so the absolute risks were calculated given that the individual has been free of breast cancer of any subtype.

Analyses were carried out in R v.3.0.2 and Stata v.14.2. All tests of statistical significance were two-sided. Further details are provided in the [Supplemental Material and Methods](#).

Results

Development of the PRS

We tried several approaches to develop PRSs; here we report results for models giving the highest prediction accuracy. Using stepwise forward selection, the best PRS for prediction of overall breast cancer was obtained at a *p* value threshold for pre-selection and stepwise regression of $p < 10^{-5}$ ([Table 1](#)). The OR per unit standard deviation (SD) for this 305-SNP PRS with overall breast cancer in

Table 1. Comparison of Methods for Deriving the PRS: Results for Overall Breast Cancer in the Validation Set

p Value Cutoff ^a	SNPs Entering Model (n)	SNPs Selected (n)	OR ^b	95% CI	AUC
Published PRS⁷					
	77	77	1.49	1.44–1.56	0.612
Hard-Thresholding Stepwise Forward Regression					
$<5 \times 10^{-8}$	1,817	123	1.59	1.52–1.66	0.626
$<10^{-6}$	2,603	197	1.62	1.55–1.68	0.634
$<10^{-5}$	3,818	305	1.65	1.58–1.72	0.637
$<10^{-4}$	6,743	669	1.62	1.56–1.69	0.631
$<10^{-3}$	14,760	1,707	1.55	1.49–1.62	0.623
Penalized Regression					
Lasso	15,032	3,820	1.71	1.64–1.79	0.647

^aThe p value cut off refers to the SNPs considered based on their marginal associations in the training set; the same p value threshold was used in each case in the stepwise regression. Parameter selection and effect size estimation for derivation of the PRS was carried out in the training set as described in the [Material and Methods](#).

^bOR per 1 SD for the PRS. OR for association with breast cancer in the validation set was derived using logistic regression adjusting for country and ten PCs. AUCs were adjusted for country. The lasso was carried out after pre-selecting SNPs at $p < 10^{-3}$ based on their marginal association in the training set. For the lasso $\lambda = 0.003$ gave the optimal PRS in the validation set.

the validation set was 1.65 (95%CI: 1.58–1.72), compared with 1.59 (95%CI: 1.52–1.66) using a “genome-wide” ($p < 5 \times 10^{-8}$) threshold (123 SNPs).

Using lasso regression, the best PRS (OR = 1.71, 95%CI: 1.64–1.79) was more predictive than the best PRS developed using the stepwise regression model. In the best model ($\lambda = 0.003$), 3,820 SNPs were selected ([Table 1](#)).

Optimizing the PRS for Prediction of Subtype-Specific Disease

For evaluation of subtype-specific models following stepwise regression, SNP effect sizes were estimated, in the first instance, in each disease subtype. The best subtype-specific PRSs using this method were also obtained at a p value threshold of $p < 10^{-5}$ ([Table S5](#)). The 305-SNP PRS was supplemented with 6 additional SNPs associated with ER-positive at p value $< 10^{-6}$ and, in addition, by two known rare breast cancer susceptibility variants in the *BRCA2* and *CHEK2* genes, bringing the total number of SNPs included to 313 (PRS₃₁₃).

The optimum subtype-specific PRS was obtained when a subset of these 313 SNPs (196 SNPs with a case-only p value for association with ER-negative versus ER-positive disease of $p < 0.025$) were given subtype-specific weights, while the remaining SNPs were given overall breast cancer weights. For ER-negative disease, the OR improved from OR = 1.45 (95%CI: 1.35–1.56) to OR = 1.47 (95%CI: 1.37–1.58) using the hybrid method compared with using only subtype-specific estimates, while for ER-positive disease the results were similar (OR = 1.74) ([Tables S6](#) and [S7](#)).

Subtype-specific prediction using the lasso analysis was optimized using case-only lasso analysis. The OR per 1 SD in the validation set was 1.81 (95%CI: 1.73–1.89) for ER-positive and 1.48 (95%CI: 1.37–1.59) for ER-negative disease ([Tables 2](#) and [S8](#)).

Validation of the PRS in the Prospective Test Dataset

The final PRSs were evaluated using data from 11,428 invasive breast cancer-affected case subjects and 18,323 control subjects from ten prospective studies. The ORs for both the overall and subtype-specific PRSs were slightly lower in the prospective test set compared to the validation set ([Table 2](#)). The difference between validation and test set may reflect some overfitting due to choosing the optimum p value threshold and for the lasso, the optimum lambda, in the validation set, but could also be due to somewhat different characteristics of the prospective studies. The ORs for overall and ER-positive, but not ER-negative, breast cancer were slightly higher for the 3,820-SNP PRS (PRS₃₈₂₀) compared with PRS₃₁₃.

The odds ratio (OR) for overall disease per 1 standard deviation (SD) of the PRS₃₁₃ in the prospective studies was 1.61 (95%CI: 1.57–1.65) while for the 77-SNP PRS (PRS₇₇) derived previously OR = 1.46 (95%CI: 1.42–1.49). For ER-negative disease the difference was OR = 1.45 (95%CI: 1.37–1.53) versus 1.35 (95%CI: 1.27–1.43) ([Table 2](#)).

The associations between the PRS and overall, ER-positive, and ER-negative breast cancer by percentiles of the PRS₃₁₃ are shown in [Figure 1](#) and [Table S9](#). Compared with women in the middle quintile (40th to 60th percentile), those in the highest 1% of risk for the subtype-specific PRS₃₁₃ had 4.37 (95%CI: 3.59–5.33)- and 2.78 (95%CI: 1.83–4.24)-fold risks, and those in the lowest 1% had 0.16 (95%CI: 0.09–0.30)- and 0.27 (95%CI: 0.09–0.86)-fold risks of developing ER-positive and ER-negative disease, respectively. The ORs by percentile of the PRS₃₈₂₀ were similar ([Table S10](#)).

Goodness of Fit of the PRS

The remaining analyses concentrated on PRS₃₁₃. The associations between the PRS and breast cancer risk by

Table 2. Association between PRS and Breast Cancer Risk in the Validation Set and Prospective Test Datasets

	Validation Set			Prospective Test Set		
	OR ^a	95% CI	AUC	OR ^a	95% CI	AUC
77 SNP PRS (PRS₇₇)						
Overall BC	1.49	1.44–1.56	0.612	1.46	1.42–1.49	0.603
ER-positive	1.56	1.49–1.63	0.623	1.52	1.48–1.56	0.615
ER-negative	1.40	1.30–1.50	0.596	1.35	1.27–1.43	0.584
313 SNP PRS (PRS₃₁₃)						
Overall BC	1.65	1.59–1.72	0.639	1.61	1.57–1.65	0.630
ER-positive	1.74	1.66–1.82	0.651	1.68	1.63–1.73	0.641
ER-negative	1.47	1.37–1.58	0.611	1.45	1.37–1.53	0.601
3,820 SNP PRS (PRS₃₈₂₀)						
Overall BC	1.71	1.64–1.79	0.646	1.66	1.61–1.70	0.636
ER-positive	1.81	1.73–1.89	0.659	1.73	1.68–1.78	0.647
ER-negative	1.48	1.37–1.59	0.611	1.44	1.36–1.53	0.600

Parameter selection and effect size estimation for derivation of the PRS was carried out in the training set as described in the [Material and Methods](#). The optimal subtype-specific PRS was obtained by carrying out case-only logistic regression and estimating effect sizes in the relevant subtype for SNPs passing a *p* value of 0.025 in case-only ordinary logistic regression (ER-positive versus ER-negative disease). OR for association with breast cancer in the validation set derived using logistic regression adjusting for country and ten PCs. AUCs were adjusted for by country. In the prospective test set, logistic regression models were adjusted for study and 15 PCs. AUCs were adjusted for by study.

^aOR per 1 SD for the PRS.

percentiles of the risk score were compared with those predicted under a simple polygenic model with the PRS considered as a continuous covariate. The effect sizes did not differ from those predicted, and in particular the estimates for the highest and lowest centile were consistent with the predicted estimates ([Table S9](#)). Further tests for goodness of fit and tail-based tests (see [Material and Methods](#)) were not statistically significant at *p* < 0.05.

There was no evidence of heterogeneity in the effect sizes among studies ([Figure 2](#)). All studies showed a significant association with similar effect sizes for overall and ER-positive breast cancer, and all but one study (FHRISK, based on only six case subjects) showed a significant effect for ER-negative breast cancer.

In the UK Biobank, the estimated hazard ratio (HR) for overall breast cancer per unit PRS (including 306 of the 313 SNPs) was HR = 1.59 (95%CI: 1.54–1.64) ([Figure 2](#)).

By way of comparison, we also evaluated a PRS based on 177 previously published susceptibility loci.^{1,2} The effect size for this PRS (OR = 1.61, 95%CI: 1.57–1.65) in the ten prospective studies was similar to the PRS₃₁₃. However, this estimated effect size is biased because the validation and test datasets used here contributed to the GWAS discovery datasets; in the UK Biobank this PRS (based on 174 of 177 available SNPs) performed worse (HR = 1.53, 95%CI: 1.48–1.58).

PRS Effects by Age

A weak decline in the OR with age was observed for ER-positive disease (*p* = 0.001, for the combined validation and

test set). There was some evidence that the decline in PRS OR was not linear, driven by a lower estimate below age 40 years ([Table S11](#), [Figure S2](#)). There was no evidence of a decline in the OR by age for ER-negative disease (*p* = 0.39).

Combined Effects of PRS and Breast Cancer Family History

The association between PRS and disease risk was observed for women with and without a family history ([Table 3](#)). However, there was some evidence that for ER-positive disease, the PRS OR was smaller in women with a family history (interaction OR = 0.91, *p* = 0.004). The log OR for family history was attenuated by 21% (1.59 to 1.44) and 12% (1.66 to 1.56) for ER-positive and ER-negative disease, respectively, after adjusting for the PRS ([Tables 3](#) and [S12](#)).

Absolute Risk of Developing Breast Cancer According to the PRS

Estimated lifetime and 10-year absolute risks for UK women in percentiles of the PRS are shown in [Figure 3](#). For ER-positive disease, the estimated lifetime absolute risk by age 80 years ranged from 2% for women in the lowest centile to 31% in the highest centile, while for ER-negative disease, the absolute risks ranged from 0.55% to 4%. The average 10-year absolute risk of breast cancer for a 47-year-old woman (i.e., the age at which women become eligible to enter the UK breast cancer screening program) in the general population is 2.6%. However, the 19% of women with the highest PRSs will attain this level of risk by age 40 years.

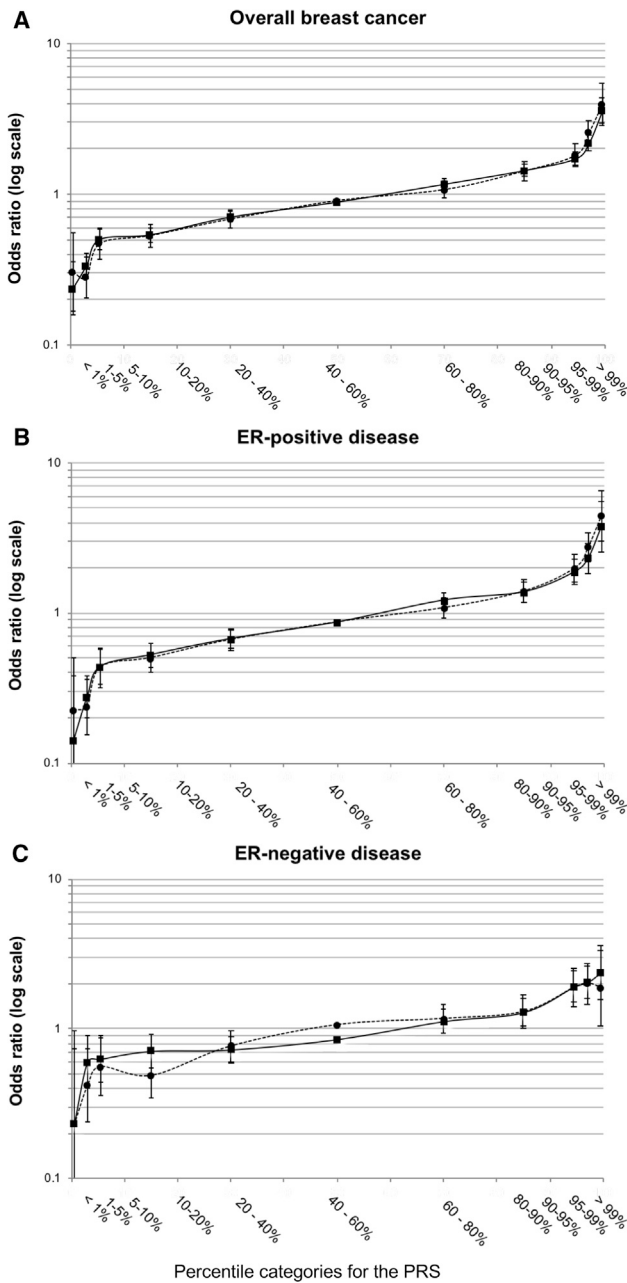


Figure 1. Association between the 313 SNP Polygenic Risk Score and Breast Cancer Risk

Association between the 313 SNP polygenic risk score (PRS) and breast cancer risk in women of European origin for (A) overall breast cancers, (B) estrogen receptor (ER)-positive disease, and (C) ER-negative disease, in the validation (dashed line) and test (solid line) sets. Odds ratios are for different quantiles of the PRS relative to the mean PRS. Odds ratios and 95% confidence intervals are shown.

Discussion

We report development and independent validation of polygenic risk scores for breast cancer, optimized for prediction of subtype-specific disease and based on the largest available GWAS dataset. The best PRS based on a hard thresholding approach included 313 SNPs and was signifi-

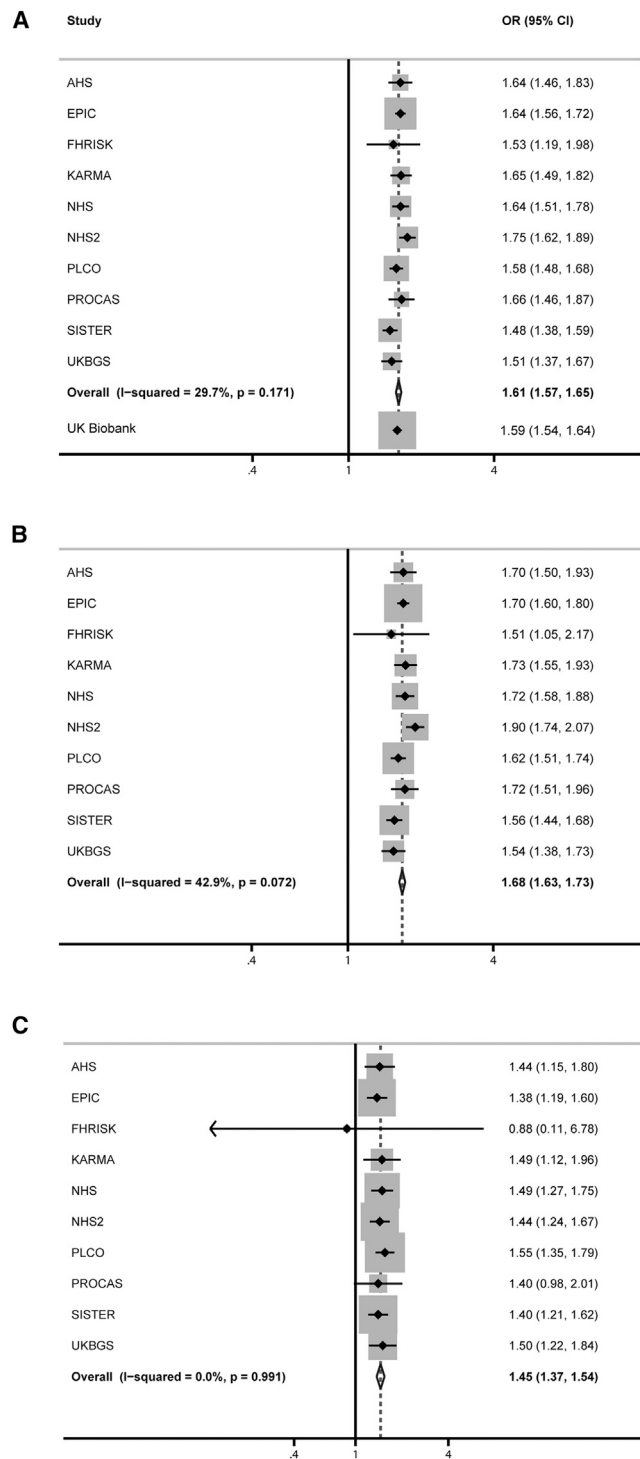


Figure 2. Prospective Validation for the 313 SNP Polygenic Risk Score

Prospective validation for the 313 SNP polygenic risk score (PRS) by study for (A) overall breast cancer, (B) ER-positive disease, and (C) ER-negative disease. Association between the 313 SNP PRS and breast cancer risk in women of European origin. Odds ratios and 95% confidence intervals are shown. I-squared and p value for heterogeneity were calculated using fixed effect meta-analysis.

cantly more predictive of risk than the previously reported 77-SNP PRS⁷ (OR per 1 SD in the prospective test set: 1.61 versus 1.46; Table 2). The effect sizes were remarkably

Table 3. Associations between the 313-SNP PRS (PRS₃₁₃) and Breast Cancer Risk by First-Degree Family History of Breast Cancer in the Combined Validation and Prospective Test Dataset

Model	ER-Positive Disease		ER-Negative Disease	
	OR ^a	95% CI	OR ^a	95% CI
Association of PRS and Breast Cancer Risk by Family History				
PRS unadjusted	1.67	1.62–1.72	1.44	1.37–1.54
PRS in women without family history	1.71	1.65–1.78	1.45	1.36–1.57
PRS in women with family history	1.55	1.48–1.65	1.40	1.27–1.55
Interaction between PRS and family history	0.91	0.85–0.97 (p = 0.004)	0.96	0.85–1.09 (p = 0.53)
Association between Family History and Breast Cancer Risk (Adjusted and Unadjusted for PRS)				
Family history unadjusted for PRS	1.59	1.46–1.72	1.66	1.41–1.95
Family history adjusted for PRS	1.44	1.33–1.57	1.56	1.32–1.83

Association with breast cancer risk was tested for using logistic regression adjusting for study and ten PCs. For these analyses the validation and test datasets were combined. Analyses were restricted to women with known age and family history information. For ER-negative disease, 4,440 women with and 13,132 women without a family history of breast cancer were included in these analyses. For ER-positive disease, 6,787 women with and 17,351 women without a family history of breast cancer were included in these analyses.

^aOR per 1 SD for the PRS.

consistent among the 10 cohorts in the prospective test set, and also consistent with that in the UK Biobank cohort (HR = 1.59, 95%CI: 1.54–1.64).

Recently, Khera et al.²⁷ derived a PRS using our publicly available summary statistics based on analysis of the BCAC data.¹ We were able to construct a PRS based on 5,194 of their 5,218 listed SNPs and compared this to our 313-SNP PRS. In our analysis of this PRS in the prospective UK Biobank data, we obtained a HR of 1.49 (95%CI: 1.44–1.54), substantially lower than that for our PRS₃₁₃. The corresponding AUCs were 0.613 (95%CI: 0.603–0.623) for their 5,194-SNP PRS versus AUC 0.630 (95%CI: 0.620–0.640) for PRS₃₁₃. Similarly, PRS₃₁₃ performed better than the Khera et al. PRS in a Biobank dataset consisting of 7,113 case subjects diagnosed before entry and 183,536 control subjects (AUC = 0.642 versus AUC = 0.627). Khera et al. report a much higher AUC (0.68), perhaps reflecting the inclusion of predictors other than SNPs in their model (for example age or principal components).

We specifically aimed to improve prediction for ER-negative breast cancer as to date prediction of this more aggressive disease has been poor. SNP selection was based on association with either ER-negative or overall breast cancer, and the optimum subtype-specific PRSs were derived by weighting a subset of SNPs according to subtype-specific effect sizes, with overall breast cancer weights used for the remaining SNPs. These results are consistent with the observation from genome-wide analyses that the heritability of ER-positive and ER-negative disease are partially correlated.² The performance of the PRS₃₁₃ in predicting ER-negative disease was considerably improved over the PRS₇₇ reported previously (OR = 1.45 versus 1.35). Nevertheless, the prediction is still better for ER-positive than ER-negative disease, reflecting the fact that ER-negative disease is more infrequent and hence the GWAS data are less powerful. The estimated heritability of ER-negative dis-

ease is similar to that of overall breast cancer,^{1,2} suggesting that more powerful ER-negative PRSs should be achievable with larger sample sizes.

The best PRS developed using lasso was more predictive for ER-positive disease but slightly less predictive for ER-negative disease in the prospective studies. Given the small differences between the models, we focused on PRS₃₁₃ since this should be more straightforward to implement in diagnostic laboratories using next generation sequencing. However, this will change with developing technology, and the cost effectiveness of using a large marker panel should be further investigated.

From a clinical viewpoint, an important consideration is the performance of the PRS in the tails of the distribution. According to the standard polygenic model, under which the effects of variants combine multiplicatively, the relationship between the PRS and the log-OR should be linear. The PRS was well calibrated at different quantiles. Even in this large study, we observed no deviation from this model, and in particular the observed risks in the highest and lowest centile were consistent with the predicted risk. The sample sizes in the extreme tails, however, were still relatively small, particularly for ER-negative disease.

While the AUC may appear modest, the predicted risk differences in the tails of the distribution are large. For the new PRS₃₁₃, the women in the top 1% of the distribution have a predicted risk that is approximately 4-fold larger than the risk in the middle quintile. The lifetime risk of overall breast cancer in the top centile of the PRSs, based on UK incidence and mortality data, was 32.6%. Women in the top centile would therefore meet the UK NICE definition of high risk (see [Web Resources](#)). In the general population, an estimated 3.6%, 12%, 21%, and 35% of all breast cancers would be expected to occur in women in the highest 1%, 5%, 10%, and 20% of the new

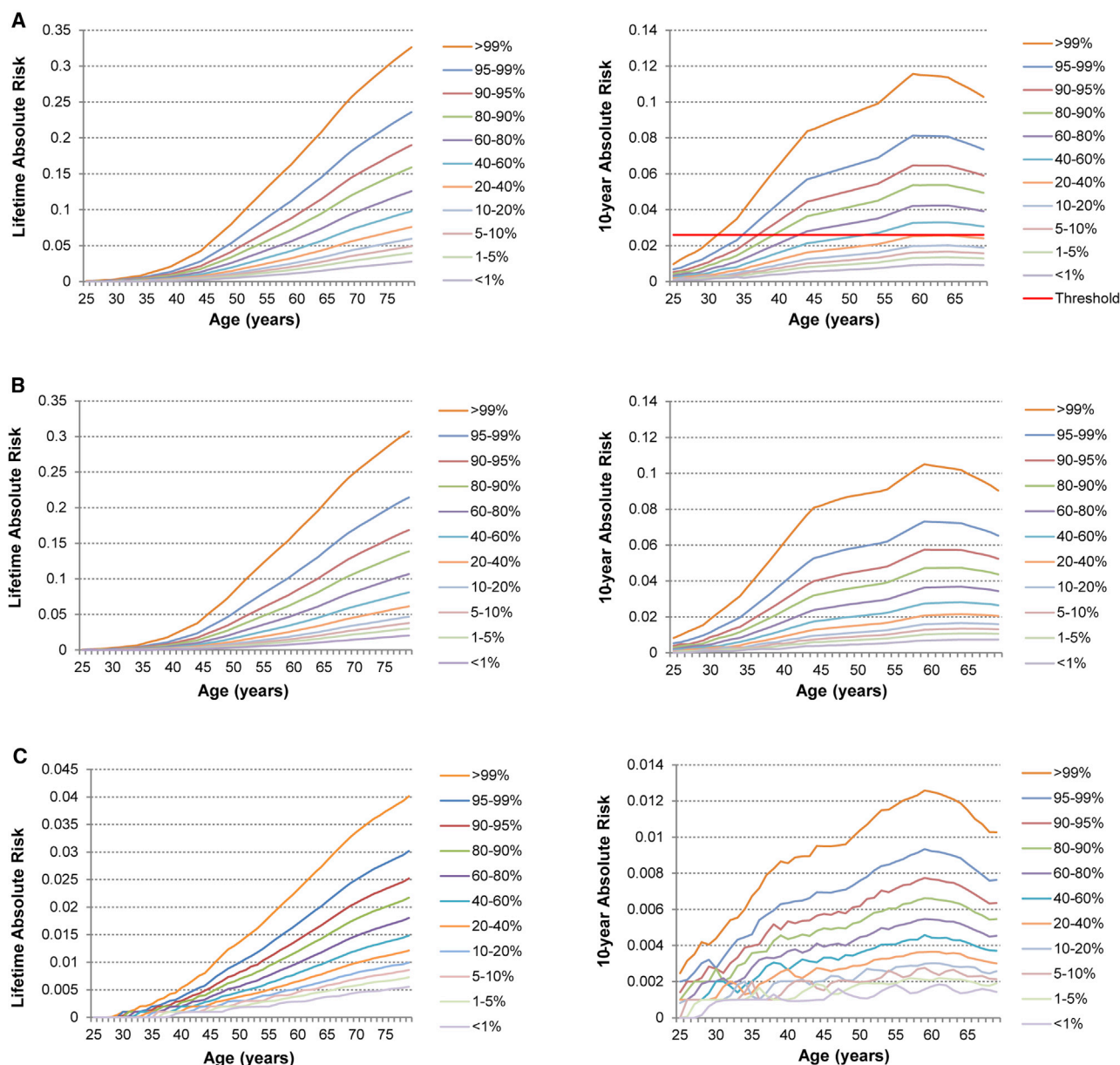


Figure 3. Cumulative and 10-Year Absolute Risk of Developing Breast Cancer

Cumulative and 10-year absolute risk of developing breast cancer for (A) overall breast cancer, (B) ER-positive disease, and (C) ER-negative disease by percentiles of the 313 SNP polygenic risk scores (PRSs). Note different scales and PRS categories in the different panels. The red line shows the 2.6% risk threshold corresponding to the mean risk for women aged 47 years. Absolute risks were calculated based on UK incidence and mortality data and using the PRS relative risks estimated as described in the [Material and Methods](#).

PRS₃₁₃, respectively, compared to only 9% of breast cancers in women in the lowest 20% of the distribution.

We observed a decline in the relative risk with age for ER-positive disease but not ER-negative disease. Even for ER-positive disease, however, the predicted relative risk, under a linear model, only declined from 1.89 at age 40 to 1.67 at age 70. While there was some indication of a lower relative risk below age 40 (estimated as 1.63 in the test set; [Figure S2](#)), these results indicate that PRS₃₁₃ is broadly applicable at all ages. We observed an attenuation of the association between breast cancer family history and breast cancer risk after adjustment for the PRS (~21% for

ER-positive, ~12% for ER-negative disease). This finding is broadly in line with the predicted contribution of the PRS to the familial relative risk of breast cancer. The PRS was predictive in women with and without a family history of breast cancer, but the OR was slightly lower in women with a family history, at least for ER-positive disease. This might reflect a weaker relative effect of the PRS in carriers of *BRCA1* or *BRCA2* mutations.²⁸ We note, however, that the absolute differences in risk by PRS will be larger in women with a family history. These results indicate that the joint effects of family history and PRS need to be considered in risk prediction.

Although we used the largest training dataset available to date for development of the PRS, further improvement should still be possible. We previously estimated using GWAS data that the theoretically best PRS, if the effect sizes of all common SNPs were known with certainty, would explain ~41% of the familial risk of breast cancer, corresponding to a standardized OR~2.1: the PRS₃₁₃ explains ~45% of this “chip” heritability.¹ This implies that larger GWASs, coupled with penalized approaches for subtype-specific disease, should further improve the predictive value of the PRS. Certain genomic features, notably transcription factor binding sites, are enriched among susceptibility loci.¹ Preliminary analyses incorporating these features into the analysis did not improve the predictive value, presumably because the enrichment effect was too small to overcome the increased complexity of the model. Better definition of genomic features to predict causal variants, and more sophisticated methods for integrating external biological information into prediction models, may improve the PRS.^{29,30}

The PRS has the potential to improve stratification for screening, while ER-specific PRSs may be informative for prevention with endocrine therapies. Previous studies have suggested that the earlier PRS₇₇ was more predictive for screen-detected breast cancers than interval cancers, and that breast cancers arising among women with a low PRS are more aggressive compared with those arising in women with a high PRS, perhaps reflecting the stronger associations with ER-positive disease.^{31,32} It will therefore be important to evaluate carefully the associations between the new PRS₃₁₃ and other tumor characteristics. Clinical translational studies are required to assess the risks and benefits of including the PRS in the context of current screening protocols.

While the PRS provides powerful risk discrimination, better risk discrimination will be obtained by combining the PRS with family history and other risk factors.¹⁰ This can be accomplished by incorporating the PRS into risk prediction models, in particular BOADICEA, which can allow for the explicit effects of family history, age, genetic, and other risk factors^{33,34} (see [Supplemental Material and Methods](#)). However, further studies to validate risk models for individualized risk prediction based on the combined effects of genetic and lifestyle risk factors will be needed. In addition, it is important to note that the PRSs generated in this study were developed and validated in white European populations and need to be validated and potentially adapted for other populations.

Accession Numbers

Requests for access to this dataset should be made to the BCAC coordinator, contact provided in [Web Resources](#).

Supplemental Data

Supplemental Data include 2 figures, 12 tables, Supplemental Acknowledgments, and Supplemental Material and Methods and

can be found with this article online at <https://doi.org/10.1016/j.ajhg.2018.11.002>.

Consortia

ABCTB Investigators are Christine Clarke, Rosemary Balleine, Robert Baxter, Stephen Braye, Jane Carpenter, Jane Dahlstrom, John Forbes, C. Soon Lee, Deborah Marsh, Adrienne Morey, Nirmala Pathmanathan, Rodney Scott, Peter Simpson, Allan Spigelman, Nicholas Wilcken, Desmond Yip, and Nikolajs Zeps.

kConFab/AOCS Investigators are Adrienne Sexton, Alex Dobrovic, Alice Christian, Alison Trainer, Allan Spigelman, Andrew Fellows, Andrew Shelling, Anna De Fazio, Anneke Blackburn, Ashley Crook, Bettina Meiser, Briony Patterson, Christine Clarke, Christobel Saunders, Clare Hunt, Clare Scott, David Amor, David Gallego Ortega, Deb Marsh, Edward Edkins, Elizabeth Salisbury, Eric Haan, Finlay Macrea, Gelareh Farshid, Geoff Lindeman, Georgia Trench, Graham Mann, Graham Giles, Grantley Gill, Heather Thorne, Ian Campbell, Ian Hickie, Liz Caldon, Ingrid Winship, James Cui, James Flanagan, James Kollias, Jane Visvader, Jennifer Stone, Jessica Taylor, Jo Burke, Jodi Saunus, John Forbes, John Hopper, Jonathan Beesley, Judy Kirk, Juliet French, Kathy Tucker, Kathy Wu, Kelly Phillips, Laura Forrest, Lara Lipton, Leslie Andrews, Lizz Lobb, Logan Walker, Maira Kentwell, Mandy Spurdle, Margaret Cummings, Margaret Gleeson, Marion Harris, Mark Jenkins, Mary Anne Young, Martin Delatycki, Mathew Wallis, Matthew Burgess, Melissa Brown, Melissa Southey, Michael Bogwitz, Michael Field, Michael Friedlander, Michael Gattas, Mona Saleh, Morteza Aghmesheh, Nick Hayward, Nick Pachter, Paul Cohen, Pascal Duijf, Paul James, Pete Simpson, Peter Fong, Phyllis Butow, Rachael Williams, Rick Kefford, Rodney Scott, Roger Milne, Rosemary Balleine, Sarah-Jane Dawson, Sheau Lok, Shona O’Connell, Sian Greening, Sophie Nightingale, Stacey Edwards, Stephen Fox, Sue-Anne McLachlan, Sunil Lakhani, Tracy Dudding, and Yolanda Antill.

NBCS collaborators are Kristine K. Sahlberg, Lars Ottestad, Rolf Kåresen, Ellen Schlichting, Marit Muri Holmen, Toril Sauer, Vilde Haakensen, Olav Engebråten, Bjørn Naume, Alexander Fosså, Cecile E. Kiserud, Kristin V. Reinertsen, Åslaug Helland, Margit Riis, Jürgen Geisler, and OSBREAC.

Acknowledgments

BCAC was funded by Cancer Research UK (C1287/A16563) and by the European Community’s Seventh Framework Programme under grant agreement no. 223175 (HEALTH-F2-2009-223175) (COGS) and by the European Union’s Horizon 2020 Research and Innovation Programme under grant agreements 633784 (B-CAST) and 634935 (BRIDGES). Genotyping of the OncoArray was principally funded by Government of Canada through Genome Canada and the Canadian Institutes of Health Research (grant GPH-129344), the Ministère de l’Économie, de la Science et de l’Innovation du Québec through Genome Québec, the Quebec Breast Cancer Foundation; NIH grants U19 CA148065 and X01HG007492; and Cancer Research UK (C1287/A10118 and C1287/A16563). Genotyping of the iCOGS array was funded by the European Union (HEALTH-F2-2009-223175), Cancer Research UK (C1287/A10710), the Canadian Institutes of Health Research for the “CIHR Team in Familial Risks of Breast Cancer” program, and the Ministry of Economic Development, Innovation and Export Trade of Quebec (grant # PSR-SIIRI-701). Combining the GWAS data was supported in part by the National Institutes

of Health (NIH) Cancer Post-Cancer GWAS initiative grant: No. 1 U19 CA 148065 (DRIVE, part of the GAME-ON initiative). We thank all the individuals who took part in these studies and all researchers, clinicians, technicians, and administrative staff who enabled this work to be carried out. For other acknowledgments and sources of funding, see [Supplemental Acknowledgments](#).

Declaration of Interests

D.G.E. reports grants from AstraZeneca and AmGen, outside the submitted work; U.M. has stock ownership and has received research funding from Abcodia Pvt Ltd.; A. Smeets reports other from MSD, outside of the submitted work; P.A.F. reports grants and personal fees from Novartis and personal fees from Pfizer, Roche, Teva, and Celgene, outside the submitted work; R.C. declares personal fees from Novartis, AstraZeneca, and Genentech, outside the submitted work. B.R. reports funding for the conduct of the clinical Success trial paid to her institution from AstraZeneca, Chugai, Lilly, Novartis, Veridex (now Janssen Diagnostics), and Sanofi Aventis. M. Robson reports grants, personal fees, and non-financial support from AstraZeneca, personal fees from McKesson, grants and personal fees from Pfizer, non-financial support from Myriad, non-financial support from Invitae, and grants from AbbVie, Tesaro, and Medivation, outside the submitted work; and M.P.L. reports personal fees from Novartis, Pfizer, Roche, Teva, AstraZeneca, Lilly, and Eisai, outside the submitted work.

Received: August 9, 2018

Accepted: November 3, 2018

Published: December 13, 2018

Web Resources

BCAC data access, <http://bcac.ccge.medschl.cam.ac.uk>

BCAC Summary statistics, <http://bcac.ccge.medschl.cam.ac.uk/bcacdata/oncoarray/gwas-icogs-and-oncoarray-summary-results/>

CORDIS, https://cordis.europa.eu/project/rcn/212694_en.html

GenomeCanada 2018 projects, https://www.genomecanada.ca/sites/default/files/2017lsarp_backgrounder_en.pdf

NICE, familial breast cancer clinical guidelines (accessed June 4, 2018), <http://guidance.nice.org.uk/CG164>

Nomis (26 March 2018), <https://www.nomisweb.co.uk/>

Office of National Statistics, <https://www.ons.gov.uk/>

West Midlands Cancer Intelligence Unit, <http://www.wmciu.nhs.uk/>

References

1. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A., et al.; NBCS Collaborators; ABCTB Investigators; and ConFab/AOCS Investigators (2017). Association analysis identifies 65 new breast cancer risk loci. *Nature* **551**, 92–94.
2. Milne, R.L., Kuchenbaecker, K.B., Michailidou, K., Beesley, J., Kar, S., Lindström, S., Hui, S., Lemaçon, A., Soucy, P., Dennis, J., et al.; ABCTB Investigators; EMBRACE; GEMO Study Collaborators; HEBON; kConFab/AOCS Investigators; and NBSC Collaborators (2017). Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer. *Nat. Genet.* **49**, 1767–1778.
3. Pashayan, N., Duffy, S.W., Chowdhury, S., Dent, T., Burton, H., Neal, D.E., Easton, D.F., Eeles, R., and Pharoah, P. (2011). Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. *Br. J. Cancer* **104**, 1656–1663.
4. Hall, P., and Easton, D. (2013). Breast cancer screening: time to target women at risk. *Br. J. Cancer* **108**, 2202–2204.
5. Burton, H., Chowdhury, S., Dent, T., Hall, A., Pashayan, N., and Pharoah, P. (2013). Public health implications from COGS and potential for risk stratification and screening. *Nat. Genet.* **45**, 349–351.
6. Torkamani, A., Wineinger, N.E., and Topol, E.J. (2018). The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* **19**, 581–590.
7. Mavaddat, N., Pharoah, P.D., Michailidou, K., Tyrer, J., Brook, M.N., Bolla, M.K., Wang, Q., Dennis, J., Dunning, A.M., Shah, M., et al. (2015). Prediction of breast cancer risk based on profiling with common genetic variants. *J. Natl. Cancer Inst.* **107**, djv036. <https://doi.org/10.1093/jnci/djv036>.
8. Maas, P., Barndahl, M., Joshi, A.D., Auer, P.L., Gaudet, M.M., Milne, R.L., Schumacher, F.R., Anderson, W.F., Check, D., Chattopadhyay, S., et al. (2016). Breast cancer risk from modifiable and nonmodifiable risk factors among white women in the United States. *JAMA Oncol.* **2**, 1295–1302.
9. Garcia-Closas, M., Gunsoy, N.B., and Chatterjee, N. (2014). Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer. *J. Natl. Cancer Inst.* **106**, dju305. <https://doi.org/10.1093/jnci/dju305>.
10. Rudolph, A., Song, M., Brook, M.N., Milne, R.L., Mavaddat, N., Michailidou, K., Bolla, M.K., Wang, Q., Dennis, J., Wilcox, A.N., et al. (2018). Joint associations of a polygenic risk score and environmental risk factors for breast cancer in the Breast Cancer Association Consortium. *Int. J. Epidemiol.* **47**, 526–536.
11. Evans, D.G., Astley, S., Stavrinou, P., Harkness, E., Donnelly, L.S., Dawe, S., Jacob, I., Harvie, M., Cuzick, J., Brentnall, A., et al. (2016). Improvement in risk prediction, early detection and prevention of breast cancer in the NHS Breast Screening Programme and family history clinics: a dual cohort study. Southampton UK: NIHR Journals Library (Programme Grants for Applied Research, No. 4.11.), <https://www.ncbi.nlm.nih.gov/books/NBK379488/doi:10.3310/pgfar04110>.
12. Shieh, Y., Eklund, M., Madlensky, L., Sawyer, S.D., Thompson, C.K., Stover Fiscalini, A., Ziv, E., Van't Veer, L.J., Esserman, L.J., Tice, J.A.; and Athena Breast Health Network Investigators (2017). Breast cancer screening in the precision medicine era: risk-based screening in a population-based trial. *J. Natl. Cancer Inst.* **109**. <https://doi.org/10.1093/jnci/djw290>.
13. Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M.J., Maranian, M.J., Bolla, M.K., Wang, Q., Shah, M., et al.; BOCS; kConFab Investigators; AOCS Group; NBCS; and GENICA Network (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat. Genet.* **47**, 373–380.
14. Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K., et al.; Breast and Ovarian Cancer Susceptibility Collaboration; Hereditary Breast and Ovarian Cancer Research Group Netherlands (HEBON); kConFab Investigators; Australian Ovarian Cancer Study Group; and GENICA (Gene Environment Interaction and Breast Cancer in Germany) Network (2013). Large-scale genotyping identifies

- 41 new loci associated with breast cancer risk. *Nat. Genet.* 45, 353–361, e1–e2.
15. Amos, C.I., Dennis, J., Wang, Z., Byun, J., Schumacher, F.R., Gayther, S.A., Casey, G., Hunter, D.J., Sellers, T.A., Gruber, S.B., et al. (2017). The OncoArray Consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomarkers Prev.* 26, 126–135.
16. Auton, A., Brooks, L.D., Durbin, R.M., Garrison, E.P., Kang, H.M., Korbel, J.O., Marchini, J.L., McCarthy, S., McVean, G.A., Abecasis, G.R.; and 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68–74.
17. O'Connell, J., Gurdasani, D., Delaneau, O., Pirastu, N., Ulivi, S., Cocca, M., Traglia, M., Huang, J., Huffman, J.E., Rudan, I., et al. (2014). A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* 10, e1004234.
18. O'Connell, J., Sharp, K., Shrine, N., Wain, L., Hall, I., Tobin, M., Zagury, J.F., Delaneau, O., and Marchini, J. (2016). Haplotype estimation for biobank-scale data sets. *Nat. Genet.* 48, 817–820.
19. Milne, R.L., Herranz, J., Michailidou, K., Dennis, J., Tyrer, J.P., Zamora, M.P., Arias-Perez, J.I., González-Neira, A., Pita, G., Alonso, M.R., et al.; kConFab Investigators; Australian Ovarian Cancer Study Group; GENICA Network; and TNBCC (2014). A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46,450 cases and 42,461 controls from the breast cancer association consortium. *Hum. Mol. Genet.* 23, 1934–1946.
20. Joshi, A.D., Lindström, S., Hüsing, A., Barrdahl, M., VanderWeele, T.J., Campa, D., Canzian, F., Gaudet, M.M., Figueroa, J.D., Baglietto, L., et al.; Breast and Prostate Cancer Cohort Consortium (BPC3) (2014). Additive interactions between susceptibility single-nucleotide polymorphisms identified in genome-wide association studies and breast cancer risk factors in the Breast and Prostate Cancer Cohort Consortium. *Am. J. Epidemiol.* 180, 1018–1027.
21. Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33, 1–22.
22. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. B* 58, 267–288.
23. Willer, C.J., Li, Y., and Abecasis, G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26, 2190–2191.
24. French, J.D., Ghousaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O'Reilly, M., Hillman, K.M., et al.; GENICA Network; and kConFab Investigators (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. *Am. J. Hum. Genet.* 92, 489–503.
25. Meyer, K.B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S.L., French, J.D., Prathalingham, R., Dennis, J., Bolla, M.K., Wang, Q., et al.; GENICA Network; kConFab Investigators; and Australian Ovarian Cancer Study Group (2013). Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. *Am. J. Hum. Genet.* 93, 1046–1060.
26. Song, M., Kraft, P., Joshi, A.D., Barrdahl, M., and Chatterjee, N. (2015). Testing calibration of risk models at extremes of disease risk. *Biostatistics* 16, 143–154.
27. Khera, A.V., Chaffin, M., Aragam, K.G., Haas, M.E., Roselli, C., Choi, S.H., Natarajan, P., Lander, E.S., Lubitz, S.A., Ellinor, P.T., and Kathiresan, S. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* 50, 1219–1224.
28. Kuchenbaecker, K.B., McGuffog, L., Barrowdale, D., Lee, A., Soucy, P., Dennis, J., Domchek, S.M., Robson, M., Spurdle, A.B., Ramus, S.J., et al. (2017). Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. *J. Natl. Cancer Inst.* 109. <https://doi.org/10.1093/jnci/djw302>.
29. Shi, J., Park, J.H., Duan, J., Berndt, S.T., Moy, W., Yu, K., Song, L., Wheeler, W., Hua, X., Silverman, D., et al.; MGS (Molecular Genetics of Schizophrenia) GWAS Consortium; GECCO (The Genetics and Epidemiology of Colorectal Cancer Consortium); GAME-ON/TRICL (Transdisciplinary Research in Cancer of the Lung) GWAS Consortium; PRACTICAL (Prostate cancer Association group To Investigate Cancer Associated aLterations) Consortium; PanScan Consortium; and GAME-ON/ELLIPSE Consortium (2016). Winner's curse correction and variable thresholding improve performance of polygenic risk modeling based on genome-wide association study summary-level data. *PLoS Genet.* 12, e1006493.
30. Pereira, M., Thompson, J.R., Weichenberger, C.X., Thomas, D.C., and Minelli, C. (2017). Inclusion of biological knowledge in a Bayesian shrinkage model for joint estimation of SNP effects. *Genet. Epidemiol.* 41, 320–331.
31. Holm, J., Li, J., Darabi, H., Eklund, M., Eriksson, M., Humphreys, K., Hall, P., and Czene, K. (2016). Associations of breast cancer risk prediction tools with tumor characteristics and metastasis. *J. Clin. Oncol.* 34, 251–258.
32. Li, J., Holm, J., Bergh, J., Eriksson, M., Darabi, H., Lindström, L.S., Törnberg, S., Hall, P., and Czene, K. (2015). Breast cancer genetic risk profile is differentially associated with interval and screen-detected breast cancers. *Ann. Oncol.* 26, 517–522.
33. Lee, A.J., Cunningham, A.P., Kuchenbaecker, K.B., Mavaddat, N., Easton, D.F., Antoniou, A.C.; Consortium of Investigators of Modifiers of BRCA1/2; and Breast Cancer Association Consortium (2014). BOADICEA breast cancer risk prediction model: updates to cancer incidences, tumour pathology and web interface. *Br. J. Cancer* 110, 535–545.
34. Macinnis, R.J., Antoniou, A.C., Eeles, R.A., Severi, G., Al Olama, A.A., McGuffog, L., Kote-Jarai, Z., Guy, M., O'Brien, L.T., Hall, A.L., et al. (2011). A risk prediction algorithm based on family history and common genetic variants: application to prostate cancer with potential clinical impact. *Genet. Epidemiol.* 35, 549–556.